Weekly Report

Junhua Lu

May 24, 2015

We had a discussion with Prof GuWT this Friday. We identified several questions.

Unfortunately, the database in the company has some bugs and Gu TY is not available on Friday. He claimed that he would fix it at next Monday.

Prof Gu think Logistic Regression is OK for the prediction problem, which is also one kind of Generalized Linear Model. And to realize this method, we need extract the data as the following data table:

ID身份证	时间（如2011.1-2015.4）	开房	网吧	本月前有无犯罪	性别	婚姻	年龄	出生地(身份证前六位)	文化程度	职业
1	2011.1									
1	2011.2									
1	2011.3									
…										
1										
1										
2	2011.1									
2	2011.2									
2	2011.3									
…										
2										
2										
…										
…										
…										

Considering the feature of Oracle database and computing capability, extraction 20000 persons records is practicable. Before extraction, we may do some statistic like knowing the proportion of criminal in the whole dataset. After that, we may use the data of 20000 persons as the training data, to train a model for prediction the condition of next month.

We also introduce several methods mentioned last week to Prof Gu. In his opinion, there are two ways to ***combine*** these methods. 1) is let each method be a *specialist*, let them predict for one person at the same time, if the majority of specialists believe the person will commit crime, then he will be considered as bad person. 2)We may assign different weights for the specialists for better results.

We also discussed about Bayesian network. Ke JM read some documents about this method, and I also read the basic ideas of BN. It may cost long time to build this network. Zhu Biao's experiment only use several hundreds and the number of attributes is small and fullBNT(matlab package) can deal with it quickly. In this model, the cause and effect(因果关系) is essential and not an easy thing to determine. We may take time to do study on it later or we may just implement it in a limited dataset. Prof Gu is not familiar with this method.

Also, since I am not familiar with Oracle database and Prof GU prefer extract the data in oracle files and process them in R directly. I will discuss this with GU TY on Monday.

Graduation thesis defence is on next week. We may experiment on several demos in lab and try our best to finish extracting the data.